DETERMINING GENDER DIFFERENTIAL ITEM FUNCTIONING FOR MATHEMATICS IN COEDUCATIONAL SCHOOL CULTURE

S. Kanageswari Suppiah Shanmugam
School of Education and Modern Languages
Universiti Utara Malaysia

kanageswari@uum.edu.my

Received: 3 March 2018 Revised: 29 September 2018 Accepted: 20 October 2018

ABSTRACT

Purpose - In an attempt to explore item characteristics that behave differently between boys and girls, this comparative study examines gender Differential Item Functioning in a school culture that is noted to be 'thriving' mathematically.

Methodology - Some 24 grade eight mathematics items from TIMSS 2003 and TIMSS 2007 released items, with equal number of computation and word problem items were administered on 460 boys and 445 girls studying in Grade Eight from three secondary Chinese-medium coeducational schools. Word problem items were defined as items set in a real-world context. Content validity was established by constructing a table of specifications. By employing the software WINSTEPS version 3.67.0 that is based on the Rasch Model for dichotomous responses, Differential Item Functioning analysis was conducted by using Mantel-Haenszel Chi-square method. DIF items were flagged when the Mantel-Haenszel probability value was less than 0.05 and classified as negligible, moderate or large DIF based on the DIF size suggested by Educational Testing Service DIF category. The focal and reference groups were girls and boys respectively. The main delimitation was substantive analysis by using expert judgment was not conducted to identify biased items.

Findings - Using Mantel-Haenszel chi-square, two moderate DIF items that assess subtraction favoured girls. They assessed *Knowing* from topics *Whole Number* and *Fraction*. Sources of DIF are linguistics density and item presentation style. The findings suggest that with only two moderate DIF mathematics items, there

is insufficient evidence to suggest that the mathematics items functioned differently between boys and girls in school culture noted for successful mathematics learning, even though linguistics complexities of the test language cannot be ignored.

Significance - While constructing Mathematics multiple choice items, careful considerations need to be given in selecting suitable numbers in composing the content of the item so that, only the correct algorithm would produce the correct answer 'if and only if' those numbers are used. The modest results of detecting two moderate DIF items nevertheless inform national testing agencies and teacher educators on the principles of building fair items as a part of their test improvement practice in the 21st Century. The novelty of this study is that gender Differential Item Functioning was studied in the context of school culture, which is notable for successful mathematics learning.

Keywords: Differential Item Functioning, coeducation, school culture, gender, computation item, word problem item, Mantel-Haenszel Chi-square, mathematics.

INTRODUCTION

Mathematics had been stereotyped as a male dominated field (Davis, 2008), even though recent findings from TIMSS and PISA international assessments display mixed results (Liu & Wilson, 2009a, 2009b; Mullis, Martin, & Foy, 2008; Mullis, Martin, Foy & Arora, 2012; Mullis, Martin, Foy, & Hooper, 2016). In the most recent TIMSS 2015 Grade Eight Mathematics, results for the 39 participating countries indicate a bigger proportion of 25 countries (Saudi, United Arab Emirates, Egypt, South Africa, Kuwait, Qatar, Turkey, Kazakhstan, Iran Islamic Rep. of, England, Malta, New Zealand, Japan, Morocco, Georgia, Korea Rep. of, Norway, United States, Australia, Israel, Slovenia, Lebanon, Lithuania, Ireland and Hong Kong SAR) recording no significant gender difference. Only one country, Chinese Taipei did not exhibit any gender differential score. A slightly higher number of seven countries recorded better performance among girls than boys (Bahrain, Botswana, Jordan, Malaysia, Oman, Singapore and Thailand) when compared to six countries that observed the reversed performance of boys performing better than girls (Canada, Chile, Italy, Sweden, Hungary and Russian Federation) (Mullis et al., 2016).

Addressing gender differential performance in mathematics seems critical as it has global implication especially since low achievement in mathematics may discourage women from pursuing a career in science and/or mathematics-related fields. Conversely, when boys perform poorly in school and drop out early, it may create gender wage gap. Some countries such as Malaysia prioritise gender differential performance as a national agenda and is concerned with "lost boys who either leave school early or with low attainment levels" (Ministry of Education Malaysia, 2013, p. E-7). Identifying the root causes of gender differences in academic performance is therefore addressing a global economic issue (Eisenkopf, Hessami, Fischbacher & Ursprung, 2012, p. 2).

> Since low achievement in mathematics may discourage women from pursuing a career in high-paying occupational fields such as engineering, it is conceivable that the inferior math performance of female students contributes to the persistence of the gender wage gap. The identification of the root causes of gender differences in academic performance is therefore a fundamental economic issue.

As revealed in the TIMSS 2015 results, finding the presence of gender differential performance in Mathematics across many countries is not uncommon. However, within the same educational system of a country, interestingly some school cultures 'thrive' mathematically due to the differences in the teaching approaches adopted by different schools. *The Telegraph*, an international daily, highlighted that for TIMSS 2015 Mathematics "East Asian countries maintaining their 20 year lead for pupils aged 10 and 14." (Gurney, 2016). In the Eastern culture, Leung (2006) further explained that Chinese or Confucian tradition school have been found to encourage procedural teaching, which is different from rote learning and that procedural learning involves repeated practices and as such, is more possible for students to understand better. In addition, their school culture also emphasise "lively learning atmosphere in class", "plenty of drills and practices", "more homework", "more tuition" as well as "more competition and quizzes" (Lim, 2003, p. 119), which explains their relatively superior results (Chan & Mousley, 2005).

In view of examining gender differential performance in the context of school culture, one alternative is to study the characteristics of

mathematics items from the perspective of item-type, which includes computation and word problem items (Gallagher, 1998; Gallagher, DeLisi, Holst, McGillicuddy-DeLisi, Morely, & Cahalan, 2000). Therefore, this study examines whether mathematics items function differently between boys and girls from coeducational secondary Chinese-medium schools, who have attended six years of schooling in primary Chinese-medium schools, whose school culture is renowned for the 'special' teaching approaches and successful mathematics learning (Lim, 2003).

Differential Item Functioning (DIF)

Differential Item Functioning (DIF) refers to the differences in item functioning *after* groups have been matched with respect to their ability. It occurs when test items function differently for students from two different comparison groups that are matched by the construct being measured (Dodeen & Johanson, 2003), which results in the probability of giving a correct response is different for both groups despite being matched based on their test proficiency. Plake and Hoover (1979) explained on the meaning of *different* as 'items that are relatively more difficult for one group than another.'

DIF analysis begins by separating all the *N* number of test items into matching subtests. In order to ensure that students' performance is comparable across items, the comparison groups are categorized according to the scores from the matching subtests at each score range. The total test scores on the matching subtest is calculated by using the relation $X = \sum_{i=0}^{n} u_i$ where u_i denotes either the score 0 (for incorrect response) or 1 (for correct response) assigned to the item *i* for each examinee. The total subtest scores is obtained by calculating $Y = \sum_{i=n+1}^{N} u_i$ (Gierl & Khaliq, 2001). Therefore, DIF analysis matches each student of equivalent ability from the focal-reference groups and makes comparisons possible across students with equal proficiency, which is very much different from t-test that compute group differences.

DIF analysis is more accurate as it is based on the observable statistical readings than the inferences that explain the nature or the

source of DIF, which is item bias (Shealy & Stout, 1993). According to Shealy and Stout (1993, p.159), test bias is a

> formalization of the intuitive idea that a test is less valid for one group of examinees than for another group and hence acts unfairly in its attempt to assess examinee differences in an intended to be measured (target) latent trait ... the process is conceptualized as individually-biased items acting in concert through a test scoring method, such as number correct, to produce a test favouring equal target ability examinees from one group over those from another.

A DIF item does not automatically guarantee that the item is biased. As Zumbo (1999, p.12) explained "DIF is a necessary condition but not a sufficient condition for item bias." This is because further analyses is needed to determine whether the item is biased or as a result of true group ability differences or impact. Impact is defined by Millsap and Everson (1993) as referring to "group differences in measured performance on tests or items." (p. 298) and involves substantive analysis, which includes judgment-based expert review. The analysis requires experts to study the structural characteristics of items such as item format or the content to locate for possible sources that influenced the probability of answering an item correctly for the different groups of students (Song, Cheng, & Klinger, 2015). Engelhard, Hansche, and Rutledge (1990) further cautioned the use of experts to determine item bias since in practice, there is a possibility of expert review failing to identify the source of DIF despite the item flagged as a DIF item. In addition, Martinková et. al., (2017, p. 3) explained that

> Even if the item flagged as DIF is later reviewed and considered fair, the act of identifying these gaps in conceptual understanding can inform teaching and, subsequently, help educators and policy makers to reduce such gaps in the future.

With criticisms lashed against a valid procedure to determine biased item, this study, therefore will only look into the statistical analyses of flagging DIF items especially when the presence of DIF items whether it is biased or fair, will eventually help to inform principles related to item development.

The Rasch Model and Mantel-Haenszel Chi-square method

Item Response Theory (IRT) is a family of mathematical models that predicts students' performance based on their ability (person ability) denoted by *q* and item characteristics such as item discrimination (*a* parameter), item difficulty (*b* parameter), pseudo-guessing (*c* parameter) and inattention (*d* parameter) (Harrison, Collins, & Müllensiefen, 2017). The four parameter IRT (4P IRT) includes all the four parameters (*a*, *b*, *c* and *d*), the 3P IRT has three parameters (*a*, *b* and *c*), the 2P IRT has two parameters (*a* and *b*) while the 1P IRT has only the *b* parameter and is occasionally referred to as the Rasch model (Hambleton, 1985), even though they have opposite philosophies of fit evaluation (Price, 2016). Elaborating further, Embretson, (1999) rationalises that the Rasch model is classified as a one-parameter model of IRT (Zhu, 1990) since both models are mathematically equal and include only the difficulty parameter as shown by Wright and Masters' equation (1982).

$$P_{ni}(x_{ni}=1|\beta_n,\delta_i)=\frac{e^{(\beta_n-\delta_i)}}{1+e^{(\beta_n-\delta_i)}}$$

where P_{ni} = the probability of student *n* with ability β_n responding correctly to item difficulty of $\delta_{i.}$. The Rasch dichotomous model used in this study is the simplest form in the family of Rasch models, which employs dichotomous scoring of correct (scored as 1) and incorrect (scored as 0) (Embretson 1999).

The Rasch model is analysed using the software Winsteps. Winsteps provides two methods for flagging DIF items, which are the Mantel-Haenszel (MH) chi-square and the Welch t-test. Theoretically, the results of the Mantel-Haenszel Chi-square method and t-test in Winsteps should be the same. Yet, in practice Mantel-Haenszel Chi-square method is found to be more accurate as it is robust to missing data, making Mantel-Haenszel Chi-square method a preferable method in detecting DIF compared to the t-test in Winsteps (Linacre, 2017). In addition, Educational Testing service being a renowned test developer uses Mantel-Haenszel Chi-square method in DIF analysis (Linacre, 2012) But in practice, Mantel-Haenszel Chi-square method will be more accurate if the data were complete. Linacre (2017) compares further the limitations of using either one DIF analysis of Mantel-Haenszel Chi-square and the t-test implemented in Winsteps

> M-H and the t-tests in Winsteps should produce the same results, because they are based on the same logit-linear theory.But, in practice, M-H will be more accurate if the data are complete and there are large numbers of subjects at every scorelevel, so called "thin" matching. Under other circumstances, M-H may not be estimable, or must use grouped-score "thick" matching, in which case the t-test method will probably be more accurate. (p. 607)

In examining further the Mantel-Haenszel Chi-square method implemented in Winsteps, Mantel-Haenszel Chi-square method does not match the reference and focal groups based on their ability level using the test scores. Instead, Winsteps transforms the raw test scores into their corresponding interval scores through the use of person measure. Therefore, the Winsteps implementation is different from the usual Mantel-Haenszel computation, which stratifies by raw scores, unlike Winsteps Mantel-Haenszel Chi-square method that stratifies the student sample by person measure (Linacre, 2017). Linacre (2017) differentiates between the conventional Mantel-Haenszel computation (M-H computation) and the Winsteps Mantel-Haenszel Chi-square method (Winsteps M-H) and cautions against treating both as same.

> The usual M-H computation stratifies the sample by raw scores, so it works with case-wise deletion of cases with missing data. Winsteps stratifies cases by measure, so cases with missing data are stratified at their estimated measure. For complete data and thin-slicing, the conventional M-H computation and the Winsteps M-H computation produce the same numbers. With missing data or thick-slicing, the conventional M-H computations and the Winsteps M-H computations may differ.(p.607)

School Culture and Mathematics DIF Items

Mathematical learning integrates the linguistic knowledge, conceptual knowledge and procedural knowledge. Linguistic knowledge involves test language proficiency, which according to Barwell (2002, p.2) involves natural language and specific mathematical

vocabulary. He further categorised specific mathematical vocabulary into three broad groups as

- technical terms specific to mathematics (e.g. equilateral, quotient, probability);
- specialist use of more general terms (e.g. line, factor, frequency);
- mathematical terms that use everyday words used for unrelated ideas (e.g. function, expression, difference, area).

Conceptual knowledge involves the understanding of the mathematical concepts, which decide the correct algorithm to perform for successful solution. Of interest is the procedural knowledge. According to the Virginia Department of Education (2004), procedural knowledge involves the different methods of teaching mathematics or the different approaches of learning mathematics that are uniquely defined by the different cultures. Therefore, school culture is an important aspect to be considered in the learning of mathematics (Lim, 2003).

According to Lim (2003), the culture of mathematics learning is 'a system of shared knowledge, practices, beliefs, and values about mathematics learning' (p. 111). With school community sharing a common system of belief, school culture naturally shapes mathematics teaching and learning. Particularly, the differences in the teaching approaches adopted by different schools that make the difference and not the students' ethnicity. Ayodele (2009) who studied on gender differential performance by school type found that the average mathematics achievement gap of boys and girls was statistically significant regardless of whether there were studying in public or private schools. His findings also revealed that there was there was a strong interaction effect between their gender, the type of school they attended and their mathematics achievement, even though the strength of the relationship was found to be stronger for Science. As such, this study is focussed on investigating gender DIF only in coeducational (mixed-gender) schools as literature on gender differential performance in single-gender and mixed-gender schools varies (Doris, O'Neill & Sweetman, 2013).

Studies on gender DIF is common and there is a large body of literature to explain gender differential performance by item-type. Among some of them, boys favour items that have figures and real-life contexts (Lane, Wang & Magone, 1996; Abedalaziz, 2010),

complex multiple choice items (Liu &Wilson, 2009a) and items of higher computational skills involving at least three mathematical operations (Salubayba, 2014). On the other hand, girls prefer routine items (Abedalaziz, 2010), word problem items (Berberoglu, 1995), computation items (Fennema & Carpenter, 1981) and conceptual explanations (Lane et al., 1996). By content domain, boys favour items from geometry while girls favour algebra (Abedalaziz, 2010). An interesting find from Abedalaziz's study (p. 113, 2010) is that boys favour items that involve basic operation such as the 'multiplication of decimal fractions' and 'abstract mathematical concepts'. His findings revealed that DIF items that favoured boys assessed the concrete mathematical concept of basic multiplication and abstract mathematical concepts.

However, DIF studies in specific school cultures, in particular that investigates the culture of 'special' teaching approaches for mathematics among students, whose dominant language is their non-English mother tongue are limited. To cite one study conducted between the Western and Eastern cultures, Liu and Wilson (2009b) who examined PISA 2003 mathematics items between students in Hong Kong and the United States of America found that the gender differential performance was more acute in the former than the latter. They also classified gender DIF items by content domains co-existing in both cultures. In both cultures, mathematics items from the content domains of geometry, and space and shape favoured boys while girls outperformed their counterpart on routine items, and items from the content domains of algebra and probability.

In attempt to add more literature about DIF items in the Eastern Asian culture and address that gap, this study aims to identify DIF items to explain gender differential performance in school culture that is stereotyped to be 'thriving mathematically'. Accordingly, the main purpose of this study is to identify mathematics items that function differently across gender groups after the students have been matched on mathematical ability and thus, examine whether test items behave differently for boys and girls with same underlying performance level, from a school culture, which is renowned for the 'special' teaching approaches for successful mathematics learning (Lim, 2003).

Therefore the research objectives are

(1) To examine the extent the data fit the Rasch model
(2) To identify DIF items that function differently between boys and girls.

(3) To investigate whether boys and girls in a school culture that has 'special' procedural learning for mathematics have different probability of producing correct responses on mathematics items, despite possessing equivalent mathematical ability.

MATERIALS AND METHODS

In this comparative study, a total of 460 boys and 445 girls in Grade Eight from three secondary Public Chinese schools sat for one-hour test that had 24 multiple choice items adopted from TIMSS 1999 and 2003 Grade Eight released items. The population had equal number of male and female students. The schools were randomly selected while cluster sampling was utilized to select the sample students. The word-problems were distinguished from computation items as items with 'real-world' setting (Neidorf, Binkley, Gattirs, & Nohara, 2006). Winsteps Version 3.67.0 (Linacre, 2008) was used for DIF analysis. The options selected by the students for each item were analysed using WINSTEPS version 3.67.0 (Linacre, 2008). The analyses conducted include determining the item mean-square (MNSQ) infit and outfit indices to determine the predictability and fit of the data. The infit and outfit MNSQ indices allow a check of the extent the data fits the Rasch model by examining the extent of the departure. The infit mean-square is affected by the examinees' response pattern to the test items, while the outfit mean-square is influenced by the examinees' unexpected responses (Wright & Linacre, 1994). After examining the fit of the model, DIF analyses was conducted to flag DIF items, and preliminary analysis to obtain an initial perspective of the characteristics of the DIF items. The focal and reference groups were girls and boys respectively.

RESULT

RQ1:Does the data fit the Rasch model

As displayed in Table 1, the average infit Mean Square is 0.99, suggesting a 1% 'deficiency in Rasch-model-predicted randomness'. By using the formula of 100*(1-0.99)/0.99 provided by Wright and Linacre (1994), there is a 1.01%, "more ambiguity in the inferred measure than modelled" by the data resembling too Guttman-like, which means that 'the item difficulty estimated from low-ability persons' does not differ "noticeably from the item difficulty estimated from high-ability persons (Wright & Linacre, 1994, p. 370). The

outfit Mean Square value of 1.02 suggests that there is a 2% of more randomness or 'noise' in the obtained data These MNSQ values are within the acceptable range of 0.5 to 1.5, and are acceptable for a good measurement (Bond & Fox, 2007; Wright & Masters, 1982). Thus, they fit the model.

The MNSQ values for both infit and outfit are also within the range of 0.8 and 1.2, which is the recommended range for high stakes multiple choice items (Wright & Linacre, 1994). The standardized score of 0.2 is neither below -2 nor above 2 and therefore the data do not indicate over predictability or under predictability. The values within the acceptable range of -1.9 to 1.9 support that the data are reasonably predictable. In addition, the Pearson correlation, represented by the raw score-to-measure correlation is -0.97, which is close to the recommended value of -1 (Linacre, 2008). The Cronbach Alpha index, which is represented by the item reliability is 0.94, suggests a high reliability (DeVellis, 2003).

Therefore, the items were productive and did not degrade the measurement, and they fit the model and demonstrate reasonable prediction. In addition, the item reliability index was 0.94, suggesting high reliability. Therefore, the data fit the model and allows further analyses to be done for the Rasch model. The standardized score, indicated by zstd was within the range of -1.9 to 1.9 suggesting that the data is reasonably predictable, with no evidence suggesting lack of predictability or over predictability.

Table 1

Summary of 24 Measured (non-Extreme) Mathematics Items

	Raw Score	Count	**Model**		**Infit**		**Outfit**	
			Measure	Error	Mean Square (MNSQ)	Zstd	MNSQ	Zstd
Mean	635.0	904.8	0.00	0.10	0.99	0.0	1.02	0.3
S.D	149.4	0.3	1.21	0.02	0.10	2.6	0.24	2.7
Max	861.0	905.0	2.58	0.17	1.24	6.7	1.48	6.7
Min	65.0	904.0	-2.45	0.08	0.80	-4.9	0.67	-4.2
real rmse 0.10		adj sd 1.21		separation 12.13		item reliability 0.09		
model rmse 0.10		adj sd 1.21		separation 12.33		item reliability 0.09		
s.e. of person mean = 0.25				item raw score-to-measure correlation = -0.98				

In Figure 1, the item to person map, which have been arranged according to student ability and item difficulty summarises item distribution across person ability. The easiest item is at the bottom and the most difficult item is located at the top of the map. The M refers to the mean of the item distribution, while S and T respectively indicate that the item mean is one standard deviation away and two standard deviations away. As displayed in Figure 1, Item C1 (*Subtract: 7003 – 4078*) is the easiest item and W23 (*A thin wire 20 centimeters long is formed into a rectangle. If the width of this rectangle is 4 centimeters, what is its length?)* is the most difficult. Item W24 (*A rectangular garden that is next to a building has a path around the other three sides. What is the area of the path?*) is the fourth easiest item, which 'students reaching the top 10% of international benchmark are likely to answer correctly' (Mullis, et al., 2000, p. 64). The distribution of all the 24 items across the sampled students suggests that more items can be added for students at the higher ability continuum as the most difficult item is W23 and it does not capture the person ability in the range of $3 \leq q \leq 4$. Furthermore, the person distribution exhibited in Figure 1 is skewed to the left, suggesting a higher proportion of students obtained high score. These findings further substantiate the 'assumption' that the students in this school culture demonstrate superior mathematics performance.

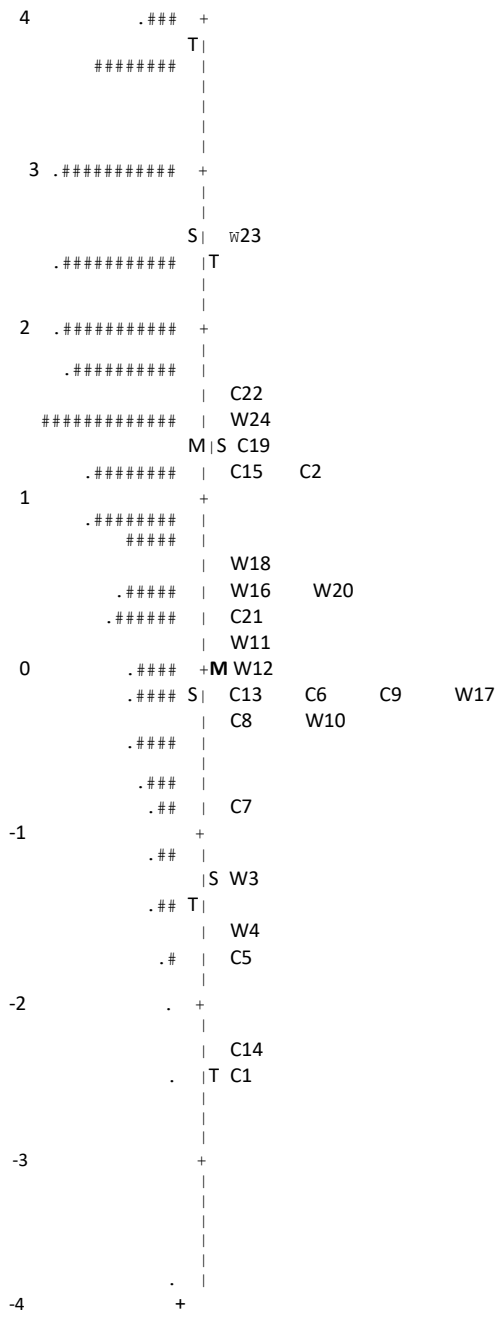

Figure 1. Item to person map.

RQ2: Which items signal negligible, moderate and adverse DIF?

Items are flagged as displaying DIF when the Mantel-Haenszel probability is less than 0.05 and categorised as negligible, moderate or large DIF based on their Mantel-Haenszel chi-square value using the Educational Testing Service (ETS) DIF category. The ETS DIF category classifies items with values more than 0.64 as displaying large DIF and values that are less than 0.43 as indicating negligible DIF. In between values in the range 0.43 to 0.64 suggest moderate DIF (Zwick et al., 1999). As exhibited in Table 2, a total of six items recorded Mantel-Haenszel probability of less than 0.05. They consist of two computation items (Items C1 and C14) and four word problem items (Items W11, W20, W23 and W24). Four of those items (Items C14, W20, W23 and W24) had Mantel-Haenszel chi-square value of less than 0.43 suggesting negligible DIF.

Only two items (C1 and W11) exhibited values between 0.43 to 0.64 and thus, were flagged as moderate DIF. One computation item (Item C1) and one word problem item (Item W11) recorded the Mantel-Haenszel chi-square size of 0.55 and 0.53 respectively, indicating slight to moderate DIF (Zwick et al., 1999). Both DIF items involve the basic operation of subtraction. Item C1 involves subtraction of two whole positive four-digit numbers, while Item W11 involves subtraction of two positive improper fraction from a whole. The positive value for the Mantel-Haenszel size indicates that the items favoured the focal group which is the girls. No large DIF items were detected**.**

Table 2

DIF Items based on IRT

Item Number	Mantel-Haenszel prob	Mantel-Haenszel Size	DIF Type	Favours
C1	0.0301	0.55	Moderate	girls
W11	0.0114	0.53	Moderate	girls
C14	0.0327	0.30	Negligible	girls
W20	0.0381	-0.27	Negligible	boys
W23	0.0079	-0.38	Negligible	boys
W24	0.0079	-0.37	Negligible	boys

RQ3: Do boys and girls in a school culture that has 'special' procedural learning for mathematics have different probability of producing correct responses on the mathematics items, despite possessing equivalent mathematical ability?

From a total of 24 items, two moderate DIF items were flagged. Item C1 (Subtract: 7003 – 4078) and Item W11 (*Ros and Jegan took cherries from a basket. Ros took* $\frac{1}{3}$ *of the cherries and Jegan took* $\frac{1}{6}$ *of the cherries. What fraction of the cherries remained in the basket?*). This suggests that when matched on mathematical ability, the boys and girls in this particular school culture recorded different probabilities of producing correct responses on a meagre 8.33 % of items. The DIF person plot for all the 24 items is illustrated in Figure 2.

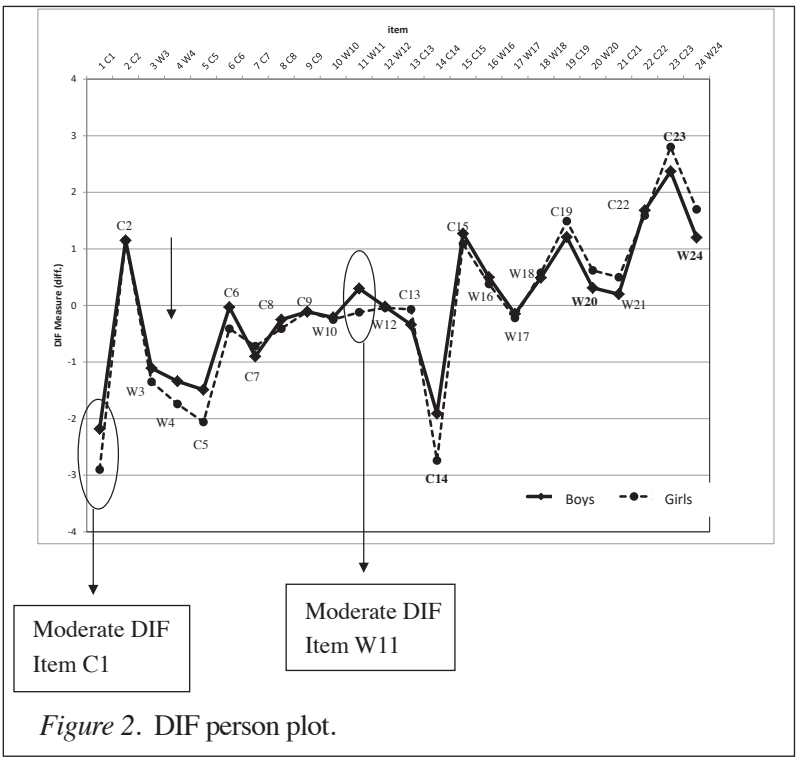

Figure 2. DIF person plot.

In examining further the two DIF items, graphical representations of the person measure were drawn to illustrate the difficulty between

the two genders. Figures 3 and 4 exhibit the person measure for the two moderate DIF items.

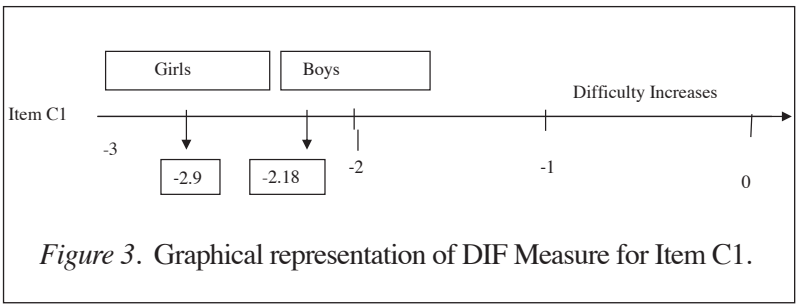

Figure 3. Graphical representation of DIF Measure for Item C1.

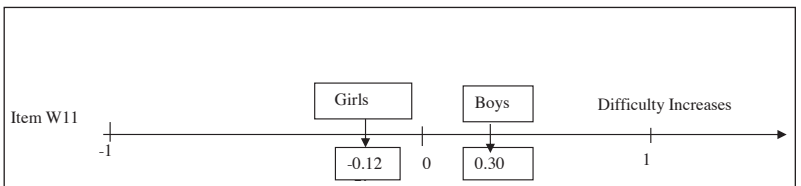

Figure 4. Graphical representation of DIF Measure for Item W11.

Comparatively, Item C1 is easier than Item W11 for the students as students with the ability level range of -3 to -2 were able to answer Item C1 than Item W11, which appealed to the student ability in the range of -0.5 to +0.5. Although these two items favoured the girls, Item C1 was easily answered by the girls who possessed relatively lower ability (q = - 2.9) than the girls with higher ability (q = - 0.12) who answered Item W11. Similarly, low ability boys (q = -2.18) were more able to answer Item C1 while high ability boys (q = 0.30) were more able to answer Item W11. These two items were also more difficult for the boys when compared to the girls.

DISCUSSION AND CONCLUSION

Item C1 is from the topic *Whole Number* and assesses the learning objective of *subtract whole numbers* of the Form One Malaysian Mathematics Curriculum. From the perspective of TIMSS, the item is from the content domain of *Fractions and Number Sense*, and assesses the cognitive domain of *using routine procedu*res. Since

Item C1 assesses a routine procedure, it is a lower-order thinking skill question (Rajendran, 2008). Connected with this one-word computation DIF item that assesses a basic arithmetic operation, the study by Abedalaziz (2010) conforms that it is possible for DIF items to occur for computing basic operation such as multiplication involving decimal fractions. The juxtaposition of his study with this present study is that his findings revealed that the DIF item assessing multiplication favoured the boys, while this study found the item assessing subtraction favouring the girls.

Delving deeper, the item only uses one word, which is *subtract* and it is a mathematical terminology that is introduced in the mathematics register of From One students. Therefore, to make claims that students did not understand the word *subtract* is remotely impossible. However, the item presentation style is atypical due to the presence of colon after the word *subtract*. In TIMSS booklet, the item was presented in the form as shown in Figure 5.

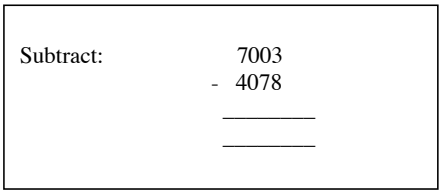

Figure 5. The original presentation of Item C1
Reprinted from Mathematics Concepts 8 Mathematics Items, by TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands, retrieved from https://nces.ed.gov/timss/pdf/TIMSS8_Math_ConceptsItems.pdf (p.30). Copyright © 2013 by International Association for the Evaluation of Educational Achievement (IEA).

The common formats of presenting arithmetic operations are in textual forms such as *Find the difference of*, *Solve,* (Choy, Kiow, Har & Hock, 2017), *Calculate* (How, Ong, Tyug, 2017) or in numerical form of *(number) – (number)* (Huat, Yeoh, How, 2016) or $\begin{array}{r} \text{number} \\ -\ \text{number} \\ \hline \end{array}$.

On the contrary, in this study this item was presented with a colon used between the word *subtract* and the numbers involved. With studies such as Gallagher (1998) and Becker (1990) that found girls favouring lower-order thinking skills question and requiring memorisation respectively than boys, it is possible that the girls applied the algorithm practised during their mathematics lesson and were not intimidated by the novel presentation style of the item. Instead, they remained faithful to the rehearsed algorithm by computing 7003 – 4078.

In examining students' solutions for Item C1, both gender groups were deselecting and selecting the correct answer. While changing options is common in any examination, one apparent conclusion derived by examining students' pattern of changing options is that there was much confusion among students. They were undecided on selecting the correct solution for Item C1, even though they knew the correct solution as they eventually selected the correct answer. Therefore, this beckons a question as to what led to such confusion to occur. Was it a simple careless mistake or bias due to the presentation style of the item, which could have been better explained if the students and the teachers were interviewed.

The other moderate DIF item, Item W11 is from the topic *Fraction* and also assesses the learning objective of subtraction, specifically *Perform subtractions involving fractions with different denominators*. From the perspective of TIMSS, this item is from the content domain of Number and assesses the cognitive domain of solving problems. This item also favoured girls and was more difficult for the boys than the girls. Unlike Item C1, Item W11 has more words. Item W11 has 28 words. Only one word is a mathematical terminology (*fraction)*, three words of two pronouns (*Ros* and *Jegan*) and the rest are natural language (*and, took, cherries, from, a, basket, of, the, what, in, the*, *remained*). Mapping against the categories outlined by Barwell (2002), *fraction* is a technical term specific to mathematics ideas, *remained* is a mathematical term that is composed of everyday words used for unrelated ideas, while the rest are natural language. The word *remained* in this context refers to *the balance left from the total*, which is synonymous to difference and indicates the operation subtraction.

One possibility is that long items require more reading when compared to shorter items and boys prefer items that involved

'less' reading, 'less writing' and that are related to their daily lives (Salubayba, 2014). Therefore, it is possible that the language inherent in the syntax influenced the linguistics complexity of the test items (Lee & Randall, 2011), especially with the presence of an everyday-word as a mathematical term used for unrelated ideas (*remained)*. Not being able to understand the mathematical meaning behind this everday-word is related to linguistics knowledge (Virginia Education Department, 2004) and would involve misunderstanding the item. The reality is students need reading comprehension skills to understand the content captured in the test items and after reading, the mathematical language embedded in the text need to be processed mathematically before writing the solutions in accordance to the required form (Kintsch & Greeno, 1985). Coming from a background where English is the students' non-mother tongue language and the Chinese language is their dominant language since it is the languages of instruction and test, and their mother tongue, the length and the linguistics nature of the item may have unnecessarily complicated further the boys who prefer 'less' reading (Salubayba, 2014).

However, if this was the case, why the other word problem items were not detected as having moderate or large DIF? Notably, a common thread binding these two items is that only these items in the studied test produced the same answer despite using incorrect algorithms. For Item C1 whether 7003 - 4078 or 4078 – 7003, the absolute value is still the same. Similarly too with Item W11, the correct solution, which is $1-\frac{1}{3}-\frac{1}{6}$ and the incorrect solution $\frac{1}{3}+\frac{1}{6}$ both produce the correct answer of $\frac{1}{2}$. Therefore, when constructing multiple choice items for Mathematics, teachers need to be aware of possible incorrect algorithms that yield the correct answer using the numbers presented as the test content. This warrants teachers to carefully select '*appropriate*' numbers as part of the test item content, so that only the correct algorithm will produce the correct answer for the learning outcome that is assessed. Taking heed of this testing principle, teachers and test developers can adopt this important implication from this study as part of their test improvement practices.

In the context of the Eastern Asian culture, specifically the Chinese-medium school culture which formed the backbone of this study, there were only 8.3% of moderate DIF items detected. Contrastive to

the study by Liu and Wilson (2009b) who discovered that Hong Kong students from the Eastern culture exhibited larger gender differential performance when compared to the American students in the Western culture, despite flagging gender DIF items favouring common content domains for both cultures. Under the present circumstances, there is a lack of strong evidence to suggest that boys and girls in a school culture that has 'special' procedural learning for mathematics have different probability of producing correct responses on mathematics items, despite possessing equivalent mathematical ability. Despite detecting two moderate DIF items that involve basic subtraction operation favouring girls, there is limited evidence to indicate items behaved differently across gender groups in school culture that is noted for successful mathematics learning, although the linguistics aspect cannot be ruled out. In conclusion, for school culture that has 'special' teaching approaches and non-English mother tongue as the instructional language, more emphasis can be invested on the structural characteristics of mathematics items (Li, Cohen & Ibarra, 2004) than on the procedural knowledge so that students are more able to 'unpack' the non-mother tongue language of the item.

Another important finding which was a limitation of this study was neither substantive analyses using expert judgement, nor interviewing students was carried out to determine the source of difficulty for both items. Therefore, conclusively identifying these two DIF items as biased was not possible. As Linacre (2008, p. 310) cautioned,

> Significance tests, such as DIF tests, are always of doubtful value in a Rasch context, because differences can be statistically significant, but far too small to have any impact on the meaning, or practical use, of the measures. So we need both statistical significance and substantive difference before we take action regarding bias.

Addressing his concern and expounding on the limitation of this study, future studies should include expert review so that some form of order can be brought to the interpretation of the sources of DIF, despite the criticism against the use of expert review to determine biased item. This is because if the source of DIF is not part of the test construct, then the biased items is a threat to test validity. The impact of DIF on test performance as Pae and Park (2006) highlighted, "can provide new insights into how DIF items in the item bank should be dealt with, and because decisions with a test are made not by the result

of an individual item score but by the result of a whole test score" (p. 476). Shedding more light on the validity of using expert review, Ercikan, Arim, Law, Domene, and Lacroix (2010) recommended an additional step of using think aloud protocol (TAP). Ercikan et al. (2010) elaborated that "evidence from expert reviews cannot be considered sufficient in deciding whether DIF items are biased and judgments about bias in test items need to include evidence from examinee thinking processes" (p. 33). This call to conduct TAP was made based on their findings from 20 biased items identified by expert review, only 10 were validated by TAP.

Therefore, identifying DIF items is as important as determining the underlying source of difficulty across the focal and reference groups. Another important recommendation for future studies is that another method of detecting DIF such as the multidimensional model to detect the presence of differential dimensions (Shelly & Stout, 1993) is used to examine whether both DIF methods flag the same items. The additional method will help to further verify the DIF items that behave differently before sources of DIF are established.

REFERENCES

Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *International Journal of Educational and Psychological Assessment, 5*, 101-116.

Ayodele, M. L. (2009). Gender differences in mathematics and integrated science achievement among junior secondary school students. *Malaysian Journal of Learning Instruction, 6*, 41-53.

Barwell, R. (2002). *The role of language in mathematics*. National Association of Language Development for curriculum. Retrieved from https://www.naldic.org.uk/Resources/NALDIC/docs/resources/documents/The%20Role%20of%20language%20in%20mathematics.pdf

Becker, B. J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. *American Educational Research Journal, 27*, 65-87.

Berberoglu, G. (1995). Differential Item Functioning (DIF) analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation, 21*(4), 439-456.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the Human Sciences, New Jersey: Lawrence Erlbaum Associates.

Chan, K.Y & Mousley, J. (2005). Using word problems in Malaysian mathematics education: Looking beneath the surface. In Chick, H. L. & Vincent, J. L. (Eds.). *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education*, 2, 217-224. Melbourne: PME.

Choy, L. C., Kiow, Y.M., Har, K.S., & Hock, L.S. (2017). *Success mathematics form 1*. Selangor, Malaysia : Oxford Fajar Sdn. Bhd.

Davis, H. (2008). Gender gaps in math and science education. *Undergraduate Research Journal for the Human Sciences*. *7*. Retrieved from http://www.kon.org/urc/v7/davis.html

DeVellis, R. (2003). *Scale development: Theory and applications*. Thousand Okas, CA: Sage.

Dodeen, H., & Johanson, G. A. (2003). An analysis of sex-related differential item functioning in attitude assessment. *Assessment & Evaluation in Higher Education, 28*(2), 129-134.

Doris, A., O'Neill, D., & Sweetman, O. (2013). Gender, single-sex schooling and maths achievement. *Economics of Education Review 35*, 104–119

Eisenkopf, G., Hessami Z., Fischbacher, U., & Ursprung, H. (2012). *Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland*, University of Konstanz Working Paper.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger (Eds.). *The new rules of measurement: What every psychologist and educator should know* (pp. 1-15). Mahwah, NJ: Lawrence Erlbaum.

Engelhard, G. Jr., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement*. *Education*. *3*(4), 347–360.

Ercikan, K., Arim, R., Law, D., Domene, J., & Lacroix, S. (2010). Application of Think Aloud Protocols for examining and confirming sources of Differential Item Functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*, 24–35.

Fennema, E., & Carpenter, T. P. (1981). Sex-related differences in mathematics: Results from National Assessment. *Mathematics Teacher, 74*, 554-559.

Gallagher, A. M. (1998) Gender and antecedents of performance in mathematics testing. *Teachers College Record*, *100*, 297-314.

Gallagher, A. M., DeLisi, R., Holst, P. C., McGillicuddy-DeLisi, A. V., Morely, M. & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*, 165-190.

Gierl, M. J. & Khaliq, S.H. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*(2), 164-187.

Gurney, J. (2016).Revealed: World pupil rankings in science and maths - TIMSS results in, The TELEGRAPH, 29 November 2016 Retrieved from http://www.telegraph.co.uk/education/2016/11/29/revealed-world-pupil-rankings-science-maths-timss-results/

Hambleton, R. K. (1985). *Item response theory*. Hingham, MA: Kluwer-Nijhoff.

Hambleton, R.K. (2006). Good practices for identifying differential item functioning. *Med Care*, *44*(3), 182–8.

Harrison, P.M.C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports, 7*(3618). 1-18. Retrieved from https://www.nature.com/articles/s41598-017-03586-z.pdf

How, N. S., Ong, C. S., & Tyug, O. Y (2017). *Mathematics Form 1*. Selangor, Malaysia: Penerbitan Pelangi Sdn. Bhd.

Huat, O. S., Yeoh, Y. K., & How, N. S. (2016). *Kurikulum standard sekolah menengah mathematik tingkatan 1*. Selangor, Malaysia: Penerbitan Pelangi Sdn. Bhd.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*(1), 109-129.

Lee, M. K., & Randall, J. (2011). Exploring language as a source of DIF in a Math test for English Language Learner. NERA Conference Proceedings. Paper 20 Retrieved from http://digitalcommons.uconn.edu/nera_2011/20

Leung, F. K. S. (2006), Mathematics education in East Asia and the West: Does culture matter?, In F. K. S. Leung, K.D. Graf &

F. J. Lopez-Real (Eds.), *Mathematics education in different cultural traditions—a comparative study of the East Asia and the West*, 21–46, New York: Springer

Neidorf, T. S., Binkley, M., Gattirs, K. & Nohara, D (2006). International Association for the Evaluation of Educational Achievement. (2013). *TIMSS 2011 Assessment*. Amsterdam, the Netherlands: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat. Retrieved from https://nces.ed.gov/timss/pdf/TIMSS2011_G8_Math.pdf

Lane, S., Wang, N. & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*(4), 21-27.

Li, Y. (2006). Using the open-source statistical language R to analyze the dichotomous Rasch model. *Behavior Research Methods, 38*(3), 532-541.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115–136.

Lim, C. S. (2003). Cultural Differences and Mathematics Learning in Malaysia. *The Mathematics Educator, 7*(1),110-122.

Linacre, J. M. (1994). *What do infit and outfit, mean square and standardized mean*. Retrieved from http://www. Rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2008). *Winstep*s (Version 3.67.0) [Computer Software]. Chicago: Winsteps.com

Linacre, J.M. (2012). *Winsteps Tutorial 4*. Retrieved from https://www.winsteps.com/a/winsteps-tutorial-4.pdf

Linacre, J. M. (2014). *Mantel and Mantel-Haenszel DIF statistics*. Retrieved from https://www.winsteps.com/winman/mantel_and_mantel-haenszel_dif.htm

Linacre, J. M. (2017). *A User's Guide to W I N S T E P S®M I N I S T E P Rasch-Model computer programs program manual 4.0.0*. Retrieved from http://www.winsteps.com/winman/copyright.htm

Liu, O.L. & Wilson, M. (2009a). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22,* 164-184.

Liu, O. L., Wilson, M. (2009b). Gender differences and similarities in PISA 2003 mathematics: A comparison between the United States and Hong Kong. *International Journal of Testing, 9,* 20-40.

Martinková P., Drabinová, A., Liaw ,Y. L., Sanders, E. A., McFarland, J. L. (2017). Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *CBE Life Science Education, 16*(2), 2-13.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334.

Mullis, I. V. S., Martin, M. O. & Foy, P. (2008). *TIMSS 2007 International Mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics Chestnut Hill*, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Mullis, I. V. S., Martin, M. O, Gonzalez, E. J., Gregory, K. D., Garden, R.A., O'Connor, K. M., Chrostowski, S.J. and Smith, T.A. (2000). *TIMSS 1999 International Mathematics Report*, International Study Center, Lynch School of Education, Boston College, Boston.

Neidorf, T. S., Binkley, M., Gattirs, K. & Nohara, D (2006). *Assessment technical report: Comparing mathematical content in the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science study (TIMSS) and the Programme for International Students Assessment (PISA) 2003*. US Education Department of Education Statistics, US Department of Education, Institute of Education Sciences, NCES 2006-029. Retrieved from http://nces.ed.gov/pubs2006/2006029-2.pdf

Waller, N. G., & Feuerstahler, L. (2017). Bayesian Modal Estimation of the Four-Parameter Item Response Model in Real,

Realistic, and Idealized Data Sets. *Multivariate Behavioral Research, 52*(3), 350-370.

Pae T., & Park G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475–496.

Plake, B. S., & Hoover, H. D. (1979). The comparability of equal raw scores obtained from in level and out-of-level testing. One source of the discrepancy between in-level and out-of level grade equivalent scores. *Journal of Educational Measurement, 16*, 271-278

Price, L. R. (2016). *Psychometric Methods: Theory into Practice*. New York, NY: Guilford.

Salubayba, T. M. (2014). Determining Differential Item Functioning in Mathematics Word Problems Using Item Response Theory. Retrieved from *http://www.iaea.info/documents/paper_226dc2c441.pdf*

Shealy, R., & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. *Papers in Language Testing and Assessment, 4*(1), 97–124.

Virginia Department of Education, (2004). *Mathematics: Strategies for teaching limited English proficient students*. Retrieved fromhttp://www.pen.k12.va.us/VDOE/Instruction/ESL/LEP math Resource.pdf

Wright, B., & Masters, G. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press

Wright, B. & Linacre (1994). *A Rasch Unidimensionality coefficient* [Electronic Version]. Retrieved from http://www.rasch.org/rmt/rmt83p.htm

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Zhu, W. (1990). *Appropriateness of the Rasch Poisson model for psychomotor test scores* (doctoral dissertation). Madison: University of Wisconsin.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now and where it is going. *Language Assessment Quarterly*, *4*(2), 223- 233.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel–Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.